DOCUMENT RESUME

ED 104 949 TM 004 407

AUTHOR Klaas, Alan C.
TITLE A Reassessment of Standard Error of Measurement.
PUB DATE [Apr 75]
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975). Document not available in hard copy due to marginal legibility of original document

EDRS PRICE MF-$0.76 HC Not Available from EDRS..PLUS POSTAGE
DESCRIPTORS *Guessing (Tests); *Scoring: *Standard Error of Measurement; Statistical Analysis; *Testing; Test Interpretation; *True Scores

ABSTRACT

Current usage and theory of standard error of measurement calls for one standard error of measurement figure to be used across all levels of scoring. The study revealed that scoring variance across scoring levels is not constant. As scoring ability increases scoring variance decreases. The assertion that low and high scoring subjects will correctly guess the same number of unknown questions is incorrect. Low and high scoring subjects correctly guess the same percentage of unknown questions. Clearly, low scoring subjects are guessing at a greater number of unknown questions and, therefore, will have a larger amount of scoring variance. Post hoc analysis of response option frequencies and changes of option choice was undertaken. Low scoring subjects apparently employ random guessing when answering questions where they do not know the correct answer. High scoring subjects seem to employ eeucated guessing when answering questions where they do not know the correct answer. Item difficulty appeared to be unrelated to the change from random guessing to educated guessing; however, the large number of omissions by low scoring subjects made the item difficulty effect difficult to interpret. (Author)

ED104949

A REASSESSMENT OF STANDARD ERROR OF MEASUREMENT

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Presented By

Dr. Alan C. Klaas
Div. of Theor. & Behav. Found.
College of Education
Wayne State University
Detroit, Michigan 48202

TM 004 407

BEST COPY AVAILABLE

To the 1975 Convention of the
American Educational Research Association

Abstract

A Reassessment of Standard Error of Measurement

Alan C. Klaas
Wayne State University

Current usage and theory of standard error of measurement calls for one standard error of measurement figure to be used across all levels of scoring. The study revealed that scoring variance across scoring levels is not constant. As scoring ability increases scoring variance decreases.

The assertion that low and high scoring subjects will correctly guess the same number of unknown questions is incorrect. Low and high scoring subjects correctly guess the same percentage of unknown questions. Clearly, low scoring subjects are guessing at a greater number of unknown questions and, therefore, will have a larger amount of scoring variance.

Post hoc analysis of response option frequencies and changes of option choice was undertaken. Low scoring subjects apparently employ random guessing when answering questions where they do not know the correct answer. High scoring subjects seem to employ educated guessing when answering questions where they do not know the correct answer. Item difficulty appeared to be unrelated to the change from random guessing to educated guessing; however, the large number of omissions by low scoring subjects made the item difficulty effect difficult to interpret.

A Reassessment of Standard Error of Measurement

Alan C. Klaas
Wayne State University

Background

The theory of standard error of measurement is part of the content of many textbooks and courses dealing with tests and measurement (Ebel, 1972; Helmsteder, 1964). Standard error theory is used by countless numbers of test publishers and test interpreters in relating to people the results of standardized tests (Henmon-Nelson, 1957; Metropolitan, 1959; Otis-Lennon, 1967). A standard error figure is often used to construct score bands or confidence intervals into which the subject's score is said to fall. The use of such score bands is undertaken so as to present as accurate as possible a picture to the subject of what his score means and does not mean.

The use of standard error of measurement by authors of textbooks, instructors in tests and measurement courses, and test score interpreters is explored in the present paper. Specifically examined is the common practice of applying a single standard error figure, the standard error at the mean, for all subjects regardless of where those subjects scored. Very low scoring subjects, middle ability subjects, and very high scoring subjects will have scores explained using the same sized scoring band or confidence interval.

Standard error of measurement explains the differences in scoring from one test administration to the next administration to be the result of random guessing at unknown answers. A subject's obtained score is said to be made up of two parts: true ability, which is the number of questions for which the subject knows the correct answers; and guessing error, which is the number of questions for which the subject does not know the correct answers but was able to guess correctly. A subject's true score is said to be made up of two parts: true score, as defined above; and perfect random guessing, which is correctly guessing the number of unknown answers equal to the probability of correct guessing randomly across all options for the item. Standard error of measurement is based upon the assumption of random guessing across all options when attempting an item for which the subject does not know the correct answer (Nunnally, 1967; McNemar, 1969).

Examples of Standard Error Applications

When an individual takes a test, he attains a certain score which is based upon the number of questions answered correctly. If that same person repeats that same test, he will usually get a different score. If the process is repeated a third time, a score different from the first two testing sessions will likely result. Different scores by the same person on the same test

create a problem to those people who would like to make some decision on the basis of a test score. A single test administration does not yield a perfectly reliable measure.

To account for inconsistent measurement of stable traits, the theory of standard error of measurement (SEM) has been developed. Standard error of measurement is used to explain why a person will score differently on successive attempts at taking the same test. There are two explanations of the SEM: the commonly used explanation and the technically accurate explanation.

The common explanation of SEM begins with a single attained score. Subject A has a raw score of 60. Perhaps the SEM figure for that test is 3 raw score points. The interpreter might explain to subject A that there is a 95% chance that the subject's true ability test score is between 54 and 66 raw score points. That is to say, the true score is somewhere in an interval about the attained score.

The technically correct explanation of SEM is just the opposite. Technically speaking, the attained scores of many administrations form an interval about the true score. Suppose subject B took the test 100 times. He would obtain 100 scores which would be formed into a frequency distribution with subject B's true score said to be the mean of that distribution. Scoring above the mean is explained as high random success of guessing at unknown items while scores below the mean would be explained as low random guessing at unknown items.

In common practice only one SEM figure is calculated for any test. Textbooks discussing SEM seem to imply, by omission, that only one SEM figure is necessary to explain inconsistent measurement. Apparently the SEM is felt to represent an equal size interval for all levels of test scoring ability. Apparently random guessing at all item response options is also felt to be equal for all levels of scoring.

The explanation of true score commonly presented is as follows. Suppose a test contains 100 items with each of the items being a four option multiple choice question. If subject C has a true ability of 60 items, then he will get at least 60 items correct. The probability of correctly guessing an unknown item in a four option question is .25. Therefore, subject C's true score would be said to equal the true ability of 60 items correct plus 40 unknown items times .25 probability of correctly guessing unknown items (60 + (40 x .25)), or a true score of 70 raw score points. Variations in attained scores are said to result from changes in the success of guessing at all options for unknown questions. Apparently, again by omission, the theory holds that random guessing at all options on unknown questions will exist regardless of scoring ability.

Methodology

A school ability test (Henmon-Nelson Test of Mental Ability Form B for grades six through nine) was administered twice to 608 eighth graders from 19 elementary schools in St. Louis, Missouri and Milwaukee, Wisconsin. The interval between administrations was three weeks. The subjects were from a wide range of socio-economic backgrounds, including a wide range of scholastic ability, and had experienced previous contact with standardized tests. The a priori data analysis procedure consisted of three steps. An additional post hoc procedure was added after initial data analysis.

First, the subjects were divided into groups based upon their scoring performance on the first administration. Subjects scoring at about three standard deviations below the mean (-3 SD) form one group, labeled -3.00. Subjects scoring at about two standard deviations below the mean (-2 SD) form the second group, labeled -2.00. The grouping process continued up the score scale to the final group, which scored at about two standard deviations above the mean (+2 SD) and was labeled +2.00. Thus the groups were formed based upon their proximity to each of the six whole number standard deviations between -3 SD and +2 SD. (The data had a skewness index of -7.17, indicating a negative skew. Examination of several papers discussing the effect of the attained skewness figure leads to the conclusion that the statistic would not be adversely affected by the observed amount of lack of normality. However, the presence of the skew would be important to avoid in future studies of this type.)

Second, a well known result of repeated administrations of the same test is the phenomenon called practice effect. The average (mean or median) score of a group from a second administration of a test will always be higher than the average score from the first administration. By subtracting the second administration score from each subject's first testing score, a change score is obtained. The change scores formed the data for the study.

Third, the statistical analysis contained two parts. The variance of the change scores for each of the subgroups was compared, using Bartlett's Test of Homoscedasticity to test if the variation of scoring were equal across all levels of scoring. The second part of the statistical analysis was a simple effects two-tailed t-test for the equality of mean change scores across all levels of scoring (Glass and Stanley, 1970).

The post hoc analysis consisted of two option response frequency summaries. The first summary was a bivariate option frequency response table, listing first testing option selected across second testing option selected, further grouped by subgroup for each item. The second summary was a simple option response

frequency and proportion table, listing first testing, second testing, and total for both tests, also grouped by subgroup for each item.

Results

Table 1 presents the variance for each subgroup for the change scores calculated by subtracting the first administration score from the second administration score. Not only were the subgroup variances found to be unequal, but the variances resulted in a definite pattern. The higher subjects scored, the lower was their variation of scoring from first to second administration.

TABLE 1

Results of Bartlett's Test of Homoscedasticity for Change Scores from Test Session One to Test Session Two

	Subgroup					
Statistic	-3	-2	-1	0	+1	+2
Variance	65.59	65.84	46.01	32.15	14.46	7.39
Number	10	34	127	204	173	23
Chi-square = 79.95*						
Degrees of Freedom = 5						

*Significant at .001

Table 2 presents the t-values and the degrees of freedom for all possible t-test comparisons of the mean change score for each subgroup. Seven statistically significant differences and eight non-significant differences were found. The location of the seven significant differences is not random, but is a definite location. All statistical analyses involving subgroups above the mean (+1.00 and +2.00) were significant, with the exception of analyses involving subgroup -3.00, which contained a small n size. The trend of the mean change scores to decrease as scoring ability increased seemed to occur in examination of mean change scores.

TABLE 2

Mean Change Scores and Related t-Values for Test of Change Scores for Test Session One to Test Session Two

Subgroup		-3	-2	-1	0	+1	+2
Mean		3.60	4.26	3.88	3.48	1.82	0.13
-2.00	t =	-0.23					
(d.f.)		(42)					
-1.00	t =	-0.12	0.28				
(d.f.)		(135)	(159)				
0.00	t =	0.06	0.70	0.58			
(d.f.)		(212)	(236)	(329)			
+1.00	t =	1.33	2.74*	3.35*	3.29*		
(d.f.)		(181)	(205)	(298)	(375)		
+2.00	t =	1.86	2.35*	2.61*	2.79*	2.05*	
(d.f.)		(31)	(55)	(148)	(225)	(194)	

*Significant in two-tailed test at .10

Three items of the 90 items were chosen as examples of option response selection made by the subjects which were grouped into subgroups of like scoring ability from within the post hoc data analysis procedure. Tables 3 and 4 present the option responding for question 15. The difficulty index for item 15 was 0.94, an easy item for the entire group of subjects. Option 3 is the correct response. Examination of the proportion of times each option was selected (Table 4) reveals that for the lowest group, -3.00, incorrect responses were randomly chosen. Subgroups -2.00 and -1.00 seemed to prefer options 1 and 2 in guessing. Subgroups +1.00 and +2.00 did not miss the item. Even for an easy item the low scoring group was responding randomly. However, middle scoring groups seemed to have an option preference. That is to say, not all of the options were equally attractive to middle scoring subjects. Small n sizes of subjects missing the item, however, dictate some caution in option preference interpretation for question 15.

TABLE 3

Subgroup Test Session One and Test Session Two Option Response Frequency for Question 15 "3, 6, 9, 12, 18, ..., ..., What two numbers should come next? (1) 19 and 20 (2) 21 and 22 (3) 21 and 24 (4) 15 and 12 (5) 17 and 16"

Subgroup (n)	Second Test Response	First Test Responses 1	2	*3	4	5	Omit
-3.00	1	1	0	0	0	0	0
	2	0	1	1	0	0	0
(10)	*3	0	1	2	1	0	0
	4	0	1	0	0	0	0
	5	0	0	1	0	1	0
	Omit	0	0	0	0	0	0
-2.00	1	3	0	1	0	0	0
	2	0	1	1	0	0	0
(34)	*3	0	1	23	0	0	0
	4	0	1	2	0	0	0
	5	0	0	1	0	0	0
	Omit	0	0	0	0	0	0
-1.00	1	0	0	3	0	0	0
	2	1	0	3	0	1	0
(127)	*3	1	5	109	0	1	0
	4	0	1	2	0	0	0
	5	0	0	0	0	0	0
	Omit	0	0	0	0	0	0
0.00	1	1	1	0	0	0	0
	2	0	1	1	0	0	0
(204)	*3	1	6	187	2	0	0
	4	0	0	3	1	0	0
	5	0	0	0	0	0	0
	Omit	0	0	0	0	0	0
+1.00	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
(173)	*3	0	1	170	0	0	0
	4	0	0	1	0	0	0
	5	0	0	0	0	0	0
	Omit	0	0	1	0	0	0
+2.00	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
(23)	*3	0	0	22	0	0	0
	4	0	0	1	0	0	0
	5	0	0	0	0	0	0
	Omit	0	0	0	0	0	0

*Correct response

TABLE 4

Option Response Summary for Question 15

Sub-group (n)	Test Session	Frequency Proportion	Response Options 1	2	*3	4	5	Omit
-3.00	First	f	1	3	4	1	1	0
		p	.10	.30	.40	.10	.10	.00
(10)	Second	f	1	2	4	1	2	0
		p	.10	.20	.40	.10	.20	.00
	Total	f	2	5	8	2	3	0
		p	.10	.25	.40	.10	.15	.00
-2.00	First	f	3	3	28	0	0	0
		p	.09	.09	.82	.00	.00	.00
(34)	Second	f	4	2	24	3	1	0
		p	.12	.06	.71	.09	.03	.00
	Total	f	7	5	52	3	1	0
		p	.10	.07	.76	.04	.01	.00
-1.00	First	f	2	6	117	0	2	0
		p	.02	.04	.92	.00	.02	.00
(127)	Second	f	3	5	116	3	0	0
		p	.02	.04	.91	.02	.00	.00
	Total	f	5	11	233	3	2	0
		p	.02	.04	.92	.01	.01	.00
0.00	First	f	2	8	191	3	0	0
		p	.01	.04	.94	.01	.00	.00
(204)	Second	f	2	2	196	4	0	0
		p	.01	.01	.96	.02	.00	.00
	Total	f	4	10	387	7	0	0
		p	.01	.02	.95	.02	.00	.00
+1.00	First	f	0	1	172	0	0	0
		p	.00	.01	.99	.00	.00	.00
(173)	Second	f	0	0	171	1	0	1
		p	.00	.00	.98	.01	.00	.01
	Total	f	0	1	343	1	0	1
		p	.00	.00	.99	.00	.00	.00
+2.00	First	f	0	0	23	0	0	0
		p	.00	.00	1.00	.00	.00	.00
(23)	Second	f	0	0	22	1	0	0
		p	.00	.00	.96	.04	.00	.00
	Total	f	0	0	45	1	0	0
		p	.00	.00	.98	.02	.00	.00

*Correct Response

Tables 5 and 6 present the option responding for question 60. The difficulty index for item 60 was 0.70, with option 3 being correct. Subgroup -3.00 subjects who answered the question incorrectly seemed to prefer option 5. However, the proportion of people in the -3.00 subgroup who omitted the item is very large, causing interpretation to be quite difficult. Subgroups -2.00 and -1.00 seemed to prefer incorrect options 1, 2, and 4. Subgroup 0.00 had an equal proportion choosing options 4 and 5, but had larger proportions preferring options 1 and 2. Subgroups above the mean, +1.00 and +2.00, seemed to have strong tendency to choose options 1 and 2. The summary of item 60 is that subjects scoring below the mean who were incorrect in responding seemed to be responding randomly. Subjects scoring above the mean who incorrectly marked the options seemed to have a strong tendency to select options 1 and 2.

Tables 7 and 8 present the option responding for question 85. The difficulty index for item 85 was 0.11 indicating a very hard item. Option 5 was the correct response. Interpretation of option selection tendencies for groups scoring below the mean is hampered because of the large number of omissions. Option selection for subgroups below the mean was felt to be random with a slight preference for option 2. The preference for option 2 was stronger in the 0.00 group with responses for the remaining option appearing to be random. The subgroups above the mean had a marked tendency to select options 2 and 5. The summary of item 85 is that low scoring subjects seemed to be selecting options randomly, and that high scoring subjects seemed to prefer two particular options.

Conclusions

Two conclusions have been drawn from the results. The first conclusion is that scoring error is not consistent across all levels of scoring. To calculate and use one SEM figure for all levels of scoring seems to be an incorrect procedure. The noted negative skew in the data creates caution before contending that the observed inverse relationship between score level and change score variance will necessarily always hold true. The study needs to be replicated several times in a test-retest setting. An interesting theoretical observation has suggested itself by the observed findings.

Standard error of measurement is usually explained in terms of a subject's random fluctuations in ability to correctly guess the answers to questions for which he has no knowledge. Some days the subject has a good day guessing and some days he has a bad day guessing. The current usage of standard error of measurement maintains that the number of unknown questions which a low scoring subject can correctly guess is equal to the number of unknown questions which a high scoring subject can correctly guess. The size of the score band or confidence interval in raw score points is uniform.

TABLE 5

Subgroup Test Session One and Test Session Two Option Response Frequency for Question 60 "The daughter of my uncle has a son. My father is her son's .... (1) cousin (2) grandfather (3) great-uncle (4) great grandfather (5) brother"

Subgroup (n)	Second Test Response	First Test Responses 1	2	*3	4	5	Omit
-3.00	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
(10)	*3	1	0	1	0	0	0
	4	0	0	0	0	0	0
	5	1	1	1	1	0	2
	Omit	0	0	0	0	0	2
-2.00	1	1	2	2	0	0	0
	2	0	2	1	0	0	2
(34)	*3	3	2	6	2	0	0
	4	0	3	3	0	0	0
	5	0	1	0	1	0	0
	Omit	0	0	0	0	0	3
-1.00	1	4	1	3	0	4	0
	2	4	4	1	3	1	6
(127)	*3	6	12	44	5	0	2
	4	5	2	8	3	0	0
	5	0	0	3	0	0	0
	Omit	0	0	2	1	0	3
0.00	1	4	1	10	1	0	0
	2	1	3	5	2	1	0
(204)	*3	10	11	125	4	3	0
	4	1	3	4	1	1	0
	5	3	0	4	3	3	0
	Omit	0	0	0	0	0	0
+1.00	1	4	0	7	0	0	0
	2	1	0	3	0	0	0
(173)	*3	4	6	139	1	1	1
	4	0	1	2	0	0	0
	5	0	0	3	0	0	0
	Omit	0	0	0	0	0	0
+2.00	1	0	0	1	0	0	0
	2	0	1	1	0	0	0
(23)	*3	0	0	19	0	0	0
	4	0	0	0	0	0	0
	5	0	0	0	0	0	0
	Omit	0	0	1	0	0	0

*Correct Response

TABLE 6

Option Response Summary for Question 60

Sub-group (n)	Test Session	Frequency Proportion	Response Options 1	2	*3	4	5	Omit
-3.00	First	f	2	1	2	1	0	4
		p	.20	.10	.20	.10	.00	.40
(10)	Second	f	0	0	2	0	6	2
		p	.00	.00	.20	.00	.60	.20
	Total	f	2	1	4	1	6	6
		p	.10	.05	.20	.05	.30	.30
-2.00	First	f	4	10	12	3	0	5
		p	.12	.29	.18	.09	.00	.15
(34)	Second	f	5	5	13	6	2	3
		p	.15	.15	.38	.18	.06	.09
	Total	f	9	15	25	9	2	8
		p	.13	.22	.37	.13	.03	.12
-1.00	First	f	19	19	61	12	5	11
		p	.15	.15	.48	.09	.04	.09
(127)	Second	f	12	19	69	18	3	6
		p	.09	.15	.54	.14	.02	.04
	Total	f	31	38	130	30	8	17
		p	.12	.15	.51	.12	.03	.07
0.00	First	f	19	18	148	11	8	0
		p	.09	.09	.73	.05	.04	.00
	Second	f	16	12	153	10	13	0
		p	.08	.06	.75	.05	.06	.00
	Total	f	35	30	301	21	21	0
		p	.08	.07	.74	.05	.05	.00
+1.00	First	f	9	7	154	1	1	1
		p	.05	.04	.89	.01	.01	.01
	Second	f	11	4	152	3	3	0
		p	.06	.02	.88	.02	.02	.00
	Total	f	20	11	306	4	4	1
		p	.06	.03	.88	.01	.01	.00
+2.00	First	f	0	1	22	0	0	0
		p	.00	.04	.96	.00	.00	.00
	Second	f	1	2	19	0	0	1
		p	.04	.09	.83	.00	.00	.04
	Total	f	1	3	41	0	0	1
		p	.02	.07	.89	.00	.00	.02

*Correct Response

TABLE 7

Subgroup Test Session One and Test Session Two Option Response Frequency for Question 85 "Circle is to ellipse as square is to: (1) oval (2) cube (3) curve (4) circle (5) diamond"

Subgroup (n)	Second Test Response	First Test Responses 1	2	3	4	*5	Omit
-3.00	1	0	1	0	1	0	0
	2	0	0	1	0	0	0
(10)	3	0	0	0	0	0	2
	4	0	0	0	0	0	0
	*5	1	0	0	1	0	0
	Omit	0	0	0	0	0	3
-2.00	1	0	0	1	0	0	0
	2	0	1	1	0	0	2
(34)	3	0	2	1	0	1	4
	4	0	0	0	0	2	5
	*5	0	0	3	1	0	1
	Omit	0	0	1	1	0	7
-1.00	1	2	3	0	1	1	8
	2	0	18	5	0	0	16
(127)	3	1	7	2	0	1	9
	4	0	0	2	0	2	4
	*5	0	3	1	2	0	2
	Omit	1	1	1	0	0	34
0.00	1	1	6	4	2	2	3
	2	2	60	5	3	5	36
(204)	3	1	11	3	2	2	9
	4	1	1	1	1	1	1
	*5	2	7	1	0	6	10
	Omit	0	1	0	0	1	13
+1.00	1	2	1	0	0	1	2
	2	8	72	6	3	10	7
(173)	3	2	3	0	1	2	0
	4	1	1	0	0	1	2
	*5	3	23	1	0	15	4
	Omit	0	0	0	0	0	2
+2.00	1	0	0	0	0	0	0
	2	0	8	0	0	0	1
(23)	3	0	0	0	0	0	0
	4	0	0	0	0	0	0
	*5	0	6	0	0	7	1
	Omit	0	0	0	0	0	0

*Correct Response

TABLE 8

Option Response Summary for Question 85

Sub-group (n)	Test Session	Frequency Proportion	Response Options 1	2	3	4	*5	Omit
-3.00	First	f	1	1	1	2	0	5
		p	.10	.10	.10	.20	.00	.50
(10)	Second	f	2	1	2	0	2	3
		p	.20	.10	.20	.00	.20	.30
	Total	f	3	2	3	2	2	8
		p	.15	.10	.15	.10	.10	.40
-2.00	First	f	0	3	7	2	3	19
		p	.00	.09	.21	.06	.09	.56
(34)	Second	f	1	4	8	7	5	9
		p	.03	.12	.24	.21	.15	.26
	Total	f	1	7	15	9	8	28
		p	.02	.10	.22	.13	.12	.41
-1.00	First	f	4	32	11	3	4	73
		p	.03	.25	.09	.02	.03	.57
(127)	Second	f	15	39	20	8	8	37
		p	.12	.31	.16	.06	.06	.29
	Total	f	19	71	31	11	12	110
		p	.07	.28	.12	.04	.05	.43
0.00	First	f	7	86	14	8	17	72
		p	.03	.42	.07	.04	.08	.35
(204)	Second	f	18	111	28	6	26	15
		p	.09	.54	.14	.03	.13	.07
	Total	f	25	197	42	14	43	87
		p	.06	.48	.10	.03	.11	.21
+1.00	First	f	16	100	7	4	29	17
		p	.09	.58	.04	.02	.17	.10
(173)	Second	f	6	106	8	5	46	2
		p	.03	.61	.05	.03	.27	.01
	Total	f	22	206	15	9	75	19
		p	.06	.60	.04	.03	.22	.05
+2.00	First	f	0	14	0	0	7	2
		p	.00	.61	.00	.00	.30	.09
(23)	Second	f	0	9	0	0	14	0
		p	.00	.39	.00	.00	.61	.00
	Total	f	0	23	0	0	21	2
		p	.00	.50	.00	.00	.46	.02

*Correct Response

Such a usage seems theoretically incorrect. It is not the number of correctly guessed unknown questions which is the same, but the percentage of correctly guessed unknown questions which is the same. Both low and high scorers have an equal probability of correctly guessing at unknown questions. However, the low scoring subject is guessing at many more unknown questions than a high scoring subject. Consequently, the low scoring subject would have greater chance for a wider variability in error of measurement than would the high scoring subject.

The second conclusion further complicates any discussion of SEM intervals. The post hoc analysis of option selection was complicated by small sample sizes in the lowest subgroup (-3.00) and by large response omissions through the middle range of scoring. The tendency of high scoring subjects to choose between two options was very pronounced. The results suggested the following statement: as scoring ability increases the process of choosing options when the item is not known appears to change from the random guessing theory (SEM) to an educated guessing theory. Figure 1 illustrates that while low scoring subjects seem to be responding randomly, high scoring subjects seem to choose most likely responses. Perhaps high scoring subjects are able to eliminate some options as being obviously wrong and then randomly choose from the remaining most likely options. Eliminating some options implies that high scorers seem to have some knowledge of the question, but not enough to answer correctly. Possessing partial knowledge and then guessing among the remaining choices is defined as educated guessing.

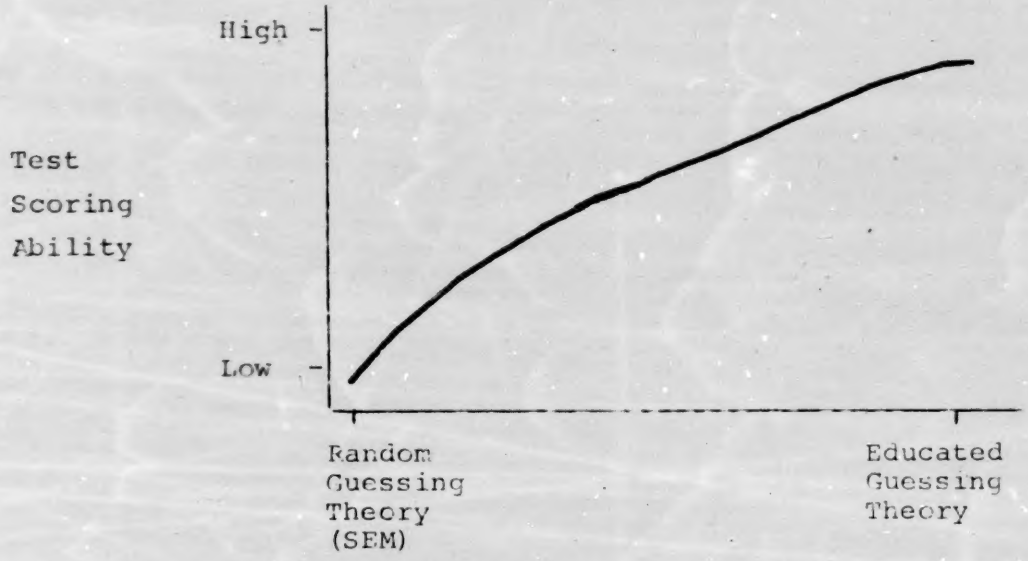

Figure 1 Illustration of Relationship Between Test Scoring Ability and Theoretical Interpretation of Guessing at Unknown Test Items.

The above conclusions need much further work with test-retest data. They are offered in hopes of stimulating such further study of an often taken-for-granted theory - standard error of measurement.

REFERENCES

Ebel, R.L. Essentials of Educational Measurement. Englewood Cliffs: Prentice-Hall, Inc., 1972.

Glass, G.V., and Stanley, J.C. Statistical Methods in Education and Psychology. Englewood Cliffs: Prentice-Hall, Inc., 1970.

Helmsteder, G.C. Principles of Psychological Measurement. New York: Appleton-Century-Crofts, 1964.

Henmon-Nelson Test of Mental Ability: Examiner's Manual. Boston: Houghton-Mifflin Co., 1957.

McNemar, Q. Psychological Statistics. New York: John Wiley and Sons, Inc., 1969.

Metropolitan Achievement Tests, Elementary Battery, Directions. New York: Harcourt, Brace, Jovanovich, 1959.

Nunnally, J.C. Psychological Theory. New York: McGraw-Hill Book Co., 1967.

Otis-Lennon Mental Abilities Test, Intermediate and Advanced Levels: Manual for Administration. New York: Harcourt, Brace, Jovanovich, 1967.